Estimation of individual genetic and environmental profiles in longitudinal designs

Boomsma, D.I.; Molenaar, P.C.M.; Dolan, C.V.

published in
Behavior Genetics
1991

DOI (link to publisher)
10.1007/BF01065818

document version
Publisher's PDF, also known as Version of record

Behavior Genetics, Vol. 21, No. 3, 1991

Estimation of Individual Genetic and Environmental Profiles in Longitudinal Designs

Dorret I. Boomsma,[1] Peter C. M. Molenaar,[2] and Conor V. Dolan[2]

Received 1 Mar. 1990—Final 22 Oct. 1990

Parameter estimates obtained in the genetic analysis of longitudinal data can be used to construct individual genetic and environmental profiles across time. Such individual profiles enable the attribution of individual phenotypic change to changes in the underlying genetic or environmental processes and may lead to practical applications in genetic counseling and epidemiology. Simulations show that individual estimates of factor scores can be reliably obtained. Decomposition of univariate, and to a lesser extent of bivariate, phenotypic time series may yield estimates of independent individual G(t) and E(t), however, that are intercorrelated. The magnitude of these correlations depends somewhat on the autocorrelation structure of the underlying series, but to obtain completely independent estimates of genetic and environmental individual profiles, at least three measured indicators are needed at each point in time.

KEY WORDS: longitudinal genetic analysis; environmental profiles; genetic profiles; factor scores; Kalman filter.

INTRODUCTION

In multivariate designs where subjects are measured on more than one variable and where correlations between variables can be explained by their loadings on one or more common factors, scores on these latent common factors can be estimated for individual subjects (Lawley and Maxwell, 1971). Estimation of factor scores can be applied in multivariate genetic modeling such as developed by Martin and Eaves (1977). In

[1] Department of Psychology, Vrije Universiteit, De Boelelaan 1115, 1081 HV Amsterdam, The Netherlands.
[2] Department of Psychology, University of Amsterdam, Roeterstraat 15, 1018 WB Amsterdam, The Netherlands.

0001-8244/91/0500-0243$06.50/0

such models multiple measures on genetically related persons provide information that is required for the estimation of individual genetic and environmental scores (Boomsma *et al.*, 1990). As more or better indicators of the latent genetic and nongenetic factors common to a set of variables are available, individual factor scores can be obtained more reliably. Estimation of individual genetic and environmental scores is numerically also possible in univariate designs, but this gives intercorrelated estimates of independent factor scores. In a univariate design, for example, MZ twins supply two observations (one on twin1 and one on twin2). Even under a simple additive genetic model, this does not provide enough information to obtain independent factor scores, since we need to estimate three factor scores (one genetic and two unique environmental scores). In longitudinal studies where the same subjects are measured repeatedly a person's genetic and environmental factor scores at each occasion can be computed yielding genetic and environmental profiles across time. As with multivariate measures taken at a single occasion, information is obtained with respect to between-subject differences in genetic and environmental deviations. High phenotypic scores in one subject, for example, may be caused by high environmental deviations and in another subject by high genetic values. In addition, a longitudinal study yields information about changes in these individual scores over time, i.e., within individuals, so that an individual increase or decrease in phenotype can be attributed to changes in the genetic or environmental contributions. In this paper we explore through simulation of longitudinal twin designs how many measures are needed at each time point to estimate reliably individual genetic and nongenetic profiles and whether the answer to this question depends on the magnitude of the genetic and environmental autocorrelations across time. To obtain these individual profiles the Kalman filter is introduced in genetic modeling.

MODEL AND METHOD

At a single time point we have the basic genetic model (discarding the subject subscript to ease presentation):

$$P_i = \lambda_i(g)G + \lambda_i(e)E + \epsilon_i, \qquad i = 1, \ldots, p, \tag{1}$$

where P is the observed phenotype, which can be univariate ($p = 1$) or multivariate ($p = 2$ or 3 in the simulations). G and E are additive genetic and environmental factor scores underlying the phenotype that are uncorrelated. The λ's are loadings of observed variables on latent factors and ϵ represents measurement error and other influences unique to each

variable and each individual. For twin data the $2p \times 2p$ covariance matrix of observations on twin1 and twin2 can be summarized as

$$\sum_P = \Lambda \Psi \Lambda' + \Theta, \tag{2}$$

where the prime indicates transpose. Λ contains the loadings on the genetic and environmental factors and is $(2p \times 4)$ since there are four latent factors ($G1$, $E1$, $G2$, $E2$), Ψ is the (4×4) correlation matrix of factor scores (e.g., Ψ contains the genetic correlations of twin1 and twin2), and θ $(2p \times 2p)$ is a diagonal covariance matrix of unique variances. Λ and θ can be estimated without the need to know the individual factor scores G and E.

Across time a quasi-Markov simplex model for the latent variables G and E can be specified. This model describes the latent G and E time series as a first-order autoregressive model (e.g., Boomsma and Molenaar, 1987; Eaves *et al.*, 1988):

$$G_{t+1} = \beta_t(g)\, G_t + \zeta_{t+1}(g) \tag{3}$$

$$E_{t+1} = \beta_t(e)\, E_t + \zeta_{t+1}(e), \tag{4}$$

where $t = 1, \ldots, T$ are the number of time points that need not be equidistant. The β's are autoregressive coefficients that describe the influence of latent factors on subsequent latent factors. The ζ_t's are innovations of the latent processes representing new genetic and environmental influences that are expressed for the first time occasion t. An illustration of this model for $p = 3$ is given in Fig. 1.

Equation (2) now can be written as

$$\sum_{PT} = \Lambda\, (I - B)^{-1}\, \Psi\, (I - B')^{-1}\, \Lambda' + \Theta, \tag{5}$$

where B $(4T \times 4T)$ has the genetic and environmental autoregression coefficients on its first lower subdiagonal and zeros elsewhere. Λ now is $(2pT \times 4T)$, Ψ is $(4T \times 4T)$, and θ is $(2pT \times 2pT)$. In regular twin designs these parameter matrices can be estimated, for example, with the LISREL VII computer program (Jöreskog and Sörbom, 1988). There are several ways to parameterize this model (see Boomsma *et al.*, 1989). In the following section we assume that Λ (factor loadings), B (autoregressive coefficients), and θ (unique variances) are estimated and that Ψ (correlations between factors) is given. Genetic and environmental factor scores can then be computed for each subject by means of Kalman filtering.

Let X_t denote the (4×1) vector of estimates of G_t and E_t at time t for twin1 and twin2 (discarding the subject subscript to ease presentation), that is, $X'_t = [G1_t, E1_t, G2_t, E2_t]$. In time-series analysis the most

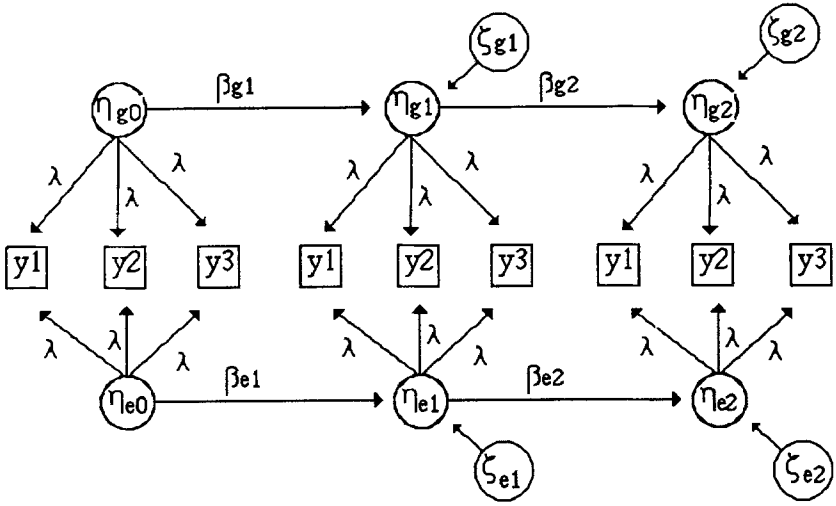

Fig. 1. Decomposition of an observed time series with three measured variables (squares) at each occasion into an underlying first-order autoregressive genetic and an underlying first-order autoregressive environmental process. λ's are factor loadings, β's are autoregressive coefficients, and ζ's are innovations of the latent processes that represent new influences entering into the process. Measurement errors have been omitted.

general approach to obtain X_t is by means of Kalman filtering (Brown, 1983) using, for example, the IMSL subroutine FTKALM (IMSL, 1979):

$$X_{t+1} = B_t X_t - K_{t+1} (\Lambda_{t+1} B_t X_t - P_{t+1}) \quad (6)$$

where B is the (4×4) matrix of autoregressive coefficients of genetic and environmental factors at time $t + 1$ on these latent factors at time t, P_{t+1} is the $(2p \times 1)$ vector of phenotypes of twin1 and twin2, and Λ $(2p \times 4)$ contains the loadings of the phenotypes at time $t + 1$ on the genetic and environmental factors. K is called the Kalman gain and is chosen in such a way that the mean square estimation error is minimized:

$$K_{t+1} = P_{t+1} \Lambda'_{t+1} (\Lambda_{t+1} P_{t+1} \Lambda'_{t+1} + \Theta_{t+1})^{-1}, \quad (7)$$

where

$$P_{t+1} = B_t V_t B_t' + \Psi_t, \quad (8)$$

where Ψ is the (4×4) correlation matrix of factor scores and V is the (4×4) covariance matrix of the sampling distribution of factor scores:

$$V_{t+1} = P_{t+1} - K_{t+1} \Lambda_{t+1} P'_{t+1} \quad (9)$$

The Kalman filter thus successively estimates X_{t+1} and V_{t+1}.

SIMULATIONS

In total 27 data sets were simulated for 100 MZ and 100 DZ twin pairs with autoregressive coefficients $\beta(g)$ and $\beta(e)$ equal to 0, 0.4, and 0.8. For each combination of $\beta(g)$ and $\beta(e)$ three data sets were simulated with one underlying genetic and with one underlying environmental series according to Eqs. (1)–(4).

1. Univariate: one indicator at each time point, $\lambda(g) = 10$ and $\lambda(e) = 10$.
2. Bivariate: two phenotypes at each time point, $\lambda_1(g) = 5$, $\lambda_2(g) = 7$, and $\lambda_1(e) = 8$, $\lambda_2(e) = 6$.
3. Trivariate: three variables at each time point, $\lambda_1(g) = 5$, $\lambda_2(g) = 7$, $\lambda_3(g) = 9$, and $\lambda_1(e) = 8$, $\lambda_2(e) = 6$, $\lambda_3(e) = 4$.

At each time point the random variables G and E are constructed to have zero means and unit variance. In the bi- and trivariate series measurement error ϵ is added to each indicator variable with var(ϵ) = 10. The uni- and bivariate series consist of 10 time points; the trivariate time series, of 8 points. Heritabilities are 0.5 at all time points in the univariate simulation series, 0.25 [i.e., 25/(25 + 64 + 10)] and 0.51 in the bivariate series, and 0.25, 0.51, and 0.76 in the trivariate series. Although in the simulations λ_i and β are the same at each occasion, this time invariance does not represent a necessary restriction and no equality constraints of this kind were used in the model fitting reported below.

RESULTS

Model Fitting

Table I gives χ^2's and probability levels that were obtained after fitting the true model to each simulated data set. For the univariate (one measured variable at each time point) and the bivariate (two measured variables at each time point) data sets, a good fit was obtained after fitting an autoregressive model consisting of a genetic and environmental simplex process to the data. For the trivariate data sets (three measured variables at each time point), however, a significant χ^2 for all nine combinations of $\beta(g)$ and $\beta(e)$ was found. The reason for this lack of fit is that the determinant of the input matrices was very small. This does not influence parameter estimates (see Table II) but biases the χ^2 upward. For the trivariate data sets the determinants of the input matrices are in the order of E–15. This situation often arises with repeated measures data. For one of the trivariate data sets [with $\beta(g) = 0.8$ and $\beta(e) = 0$]

Table I. Chi-Square Statistics and Probability Levels for Simulated Data Sets

	$\beta(g)$	$\beta(e)$	Univariate ($df=182$)		Bivariate ($df=780$)		Trivariate ($df=1035$)	
1	0	0	190.37	(.320)	805.28	(.266)[a]	1298.09	(.001)
2	0.8	0	171.07	(.709)	816.05	(.187)[a]	1314.39	(.000)
3	0	0.8	187.70	(.370)	795.30	(.344)	1256.92	(.006)
4	0.8	0.8	182.41	(.477)	780.76	(.486)	1286.18	(.001)
5	0.4	0	179.89	(.530)	806.99	(.252)[a]	1309.48	(.000)
6	0	0.4	194.59	(.248)	810.11	(.221)	1303.82	(.000)
7	0.4	0.4	188.89	(.348)	804.92	(.261)	1317.13	(.000)
8	0.4	0.8	183.47	(.456)	784.27	(.450)	1270.14	(.003)
9	0.8	0.4	177.08	(.589)	805.91	(.253)	1327.74	(.000)

[a] Measurement errors for $P1$ and $P2$ constrained to be equal. When $\beta(e)$ equals zero, no distinction can be made between $E1$ influences and measurement error otherwise.

Table II. Recovered Parameters, Trivariate Data

	True, all t	Recovered $t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
$\beta(g)$	0.8	—	.831	.856	.807	.822	.860	.827	.838
$\beta(e)$	0.0	—	.114	−.098	−.090	−.025	−.036	−.033	−.016
$\lambda1(g)$	5.0	4.93	5.29	3.90	6.21	5.25	6.07	5.74	5.47
$\lambda2(g)$	7.0	6.43	7.03	6.24	8.33	7.63	8.16	7.77	7.42
$\lambda3(g)$	9.0	8.99	8.83	8.18	10.06	9.53	10.10	9.65	9.66
$\lambda1(e)$	8.0	8.42	8.39	7.92	7.89	8.09	8.29	8.13	7.31
$\lambda2(e)$	6.0	6.54	6.19	6.23	6.21	6.16	6.18	6.21	5.53
$\lambda3(e)$	4.0	4.16	3.82	4.11	3.49	3.75	3.74	4.03	3.60
var($\epsilon1$)	10.0	10.25	10.25	10.25	10.25	10.25	10.25	10.25	10.25
var($\epsilon2$)	10.0	9.62	9.62	9.62	9.62	9.62	9.62	9.62	9.62
var($\epsilon3$)	10.0	10.17	10.17	10.17	10.17	10.17	10.17	10.17	10.17

Table II gives the recovered parameters. In spite of the significant χ^2, these are close to the values that were used to simulate the data.

Factor Scores

Estimates of Λ, B, and θ were employed in Kalman filtering of the observations to obtain genetic and environmental factor scores at each time point for each subject according to Eqs. (6)–(9). Next the estimated series were compared with the true (i.e., simulated) individual series. Table III gives correlations between true and estimated factor scores for MZ and DZ twins separately. For each data set these correlations have

Table III. Correlations of True and Estimated Factor Scores (MZ and DZ) Averaged Across Time Points for Each Data Set

			Univariate		Bivariate		Trivariate	
	$\beta(g)$	$\beta(e)$	*G*	*E*	*G*	*E*	*G*	*E*
MZ								
1	0	0	.80	.80	.84	.87	.95	.91
2	0.8	0	.84	.83	.88	.89	.96	.92
3	0	0.8	.83	.85	.87	.90	.96	.94
4	0.8	0.8	.79	.80	.89	.90	.97	.94
5	0.4	0	.82	.80	.85	.88	.96	.91
6	0	0.4	.80	.81	.85	.88	.95	.92
7	0.4	0.4	.80	.80	.86	.88	.96	.92
8	0.4	0.8	.82	.82	.87	.90	.96	.94
9	0.8	0.4	.81	.81	.88	.89	.96	.92
DZ								
1	0	0	.74	.73	.78	.80	.91	.88
2	0.8	0	.80	.78	.83	.83	.93	.89
3	0	0.8	.79	.79	.81	.86	.92	.91
4	0.8	0.8	.73	.73	.82	.86	.93	.91
5	0.4	0	.75	.73	.79	.81	.91	.88
6	0	0.4	.74	.75	.78	.82	.91	.89
7	0.4	0.4	.73	.73	.79	.82	.91	.89
8	0.4	0.8	.75	.76	.81	.85	.92	.91
9	0.8	0.4	.76	.75	.81	.83	.93	.89

been averaged across time points, as there was no difference in the correlations for earlier and later time points. It is clear that estimated *G(t)* and *E(t)* are valid indicators of the associated true scores. Correlations are higher for MZ than for DZ twins and increase as the number of measures taken at each time point increases. Even when only one measure is available, however, they are quite high already. As the number of variables at each time point increases, standard errors become smaller for both MZ and DZ twin groups, and for all data sets standard errors of estimated factor scores are smaller for MZ than for DZ twins. Table IV shows the genetic and environmental autocorrelation matrices based on the factor scores calculated in a univariate and a trivariate data set [where $\beta(g) = 0.8$ and $\beta(e) = 0$]. Even in the univariate case, the autocorrelation structures for the genetic and environmental series are correctly recovered and the dynamic structure of the true *G(t)* and *E(t)* series is accurately reflected by the individual trajectories. Finally, Table V gives the instantaneous cross-correlation between individual estimates of *G* and *E* (correlations are again averaged across time points). The expected value of this cross-correlation is zero. But for the uni- and bivariate time series this correlation may deviate from zero, because the

Table IV. Recovered Autocorrelation Matrices Based on Individual Estimates of $G(t)$ and $E(t)$: MZ Twins (Lower Half) and DZ Twins (Upper Half)

Univariate: One indicator, $\beta(g)=0.8$, $\beta(e)=0$

	*G*1	*G*2	*G*3	*G*4	*G*5	*G*6	*G*7	*G*8	*G*9	*G*10
*G*1	—	.76	.63	.52	.45	.34	.23	.21	.18	.12
*G*2	.76	—	.79	.71	.58	.49	.38	.31	.25	.18
*G*3	.59	.79	—	.83	.70	.59	.49	.40	.36	.25
*G*4	.45	.63	.83	—	.86	.75	.53	.48	.41	.26
*G*5	.40	.53	.72	.86	—	.86	.62	.56	.47	.33
*G*6	.32	.44	.55	.72	.85	—	.73	.62	.55	.47
*G*7	.09	.19	.34	.47	.59	.73	—	.84	.76	.65
*G*8	.11	.17	.30	.40	.49	.61	.86	—	.89	.71
*G*9	.04	.10	.25	.35	.45	.59	.76	.89	—	.84
*G*10	.04	.09	.18	.29	.38	.46	.58	.74	.83	—
	*E*1	*E*2	*E*3	*E*4	*E*5	*E*6	*E*7	*E*8	*E*9	*E*10
*E*1	—	.09	−.07	.01	−.05	−.04	−.09	.04	−.08	−.05
*E*2	−.02	—	−.01	.08	.01	.14	.05	−.00	−.00	−.01
*E*3	−.07	.00	—	.05	.01	−.03	.02	−.04	.03	−.03
*E*4	.08	−.08	.09	—	.12	.02	−.19	.15	−.10	−.18
*E*5	−.03	.03	.03	.05	—	.05	.04	.05	.01	.07
*E*6	.03	.11	−.08	.09	.13	—	.21	−.03	.02	.07
*E*7	−.07	.02	.10	−.05	.13	.17	—	.04	.08	.14
*E*8	−.06	.00	.04	.07	.04	−.10	.15	—	−.11	−.22
*E*9	.09	−.03	.06	.04	.05	.18	−.02	−.14	—	.12
*E*10	.01	.10	.02	.04	.05	.01	.07	.05	.07	—

Trivariate: Three indicators, $\beta(g)=0.8$, $\beta(e)=0$

	*G*1	*G*2	*G*3	*G*4	*G*5	*G*6	*G*7	*G*8
*G*1	—	.82	.68	.55	.50	.39	.26	.30
*G*2	.83	—	.85	.67	.57	.45	.25	.28
*G*3	.72	.87	—	.77	.65	.53	.34	.35
*G*4	.64	.73	.83	—	.85	.69	.55	.52
*G*5	.57	.59	.63	.81	—	.87	.69	.66
*G*6	.42	.45	.50	.65	.84	—	.83	.73
*G*7	.45	.46	.49	.63	.72	.83	—	.84
*G*8	.37	.32	.38	.47	.55	.71	.85	—
	*E*1	*E*2	*E*3	*E*4	*E*5	*E*6	*E*7	*E*8
*E*1	—	.09	.10	.04	.10	.08	.06	.02
*E*2	.14	—	−.09	−.01	.07	−.13	−.04	.01
*E*3	.04	−.09	—	−.09	.00	−.01	.03	−.06
*E*4	.15	.02	−.08	—	.01	−.02	−.09	−.04
*E*5	.13	.03	−.07	−.05	—	.04	−.09	.03
*E*6	.03	−.04	.02	.04	.02	—	−.03	−.03
*E*7	.01	−.05	−.12	−.02	−.03	−.03	—	−.01
*E*8	.02	.04	.01	−.05	.14	.15	−.02	—

Table V. Instantaneous Correlation Between Estimated G and Estimated E (Averaged Across Time Points for Each Data Set)

			Univariate		Bivariate		Trivariate	
	$\beta(g)$	$\beta(e)$	MZ	DZ	MZ	DZ	MZ	DZ
1	0	0	0.51	0.87	0.29	0.50	0.13	0.18
2	0.8	0	0.35	0.58	0.21	0.33	0.10	0.15
3	0	0.8	0.38	0.61	0.21	0.35	0.11	0.16
4	0.8	0.8	0.49	0.80	0.21	0.34	0.12	0.13
5	0.4	0	0.48	0.82	0.27	0.46	0.12	0.18
6	0	0.4	0.48	0.80	0.27	0.47	0.13	0.19
7	0.4	0.4	0.50	0.86	0.27	0.47	0.12	0.18
8	0.4	0.8	0.45	0.75	0.22	0.38	0.13	0.16
9	0.8	0.4	0.42	0.71	0.23	0.37	0.10	0.15

phenotypic series is decomposed into two trajectories. The results in Table V indicate that even in the univariate case, the correlation for MZ twins between individual estimates of G and E is consistently less than 0.5, and for the bivariate time series it does not exceed 0.3. For DZ twins on the other hand, it is more difficult in the univariate case to obtain independent estimates of genetic and environmental profiles. But in the bivariate series, the instantaneous correlation between estimates of G and E factor scores also does not exceed 0.5. The correlation between individual estimates of G and E depends somewhat on the pattern of the autocorrelations: when the difference between $\beta(g)$ and $\beta(e)$ is large, the individual decomposition yields factor scores that have a lower intercorrelation than when $\beta(g)$ and $\beta(e)$ are the same. Notice that these results pertain only to the analyses of individual estimates of $G(t)$ and $E(t)$ and do not relate in any way to the standard applications of the simplex model in genetic modeling. In the latter, only a decomposition of the total population variance is considered, whereas these results concern the relationship of individual estimates of G and E factor scores.

Finally, Fig. 2 gives the recovered genetic profiles for a pair of DZ twins and the confidence intervals around the recovered scores. These recovered profiles closely follow the true profiles. For this example, data from the last trivariate data set were used [where $\beta(g) = 0.8$ and $\beta(e) = 0.4$]. The confidence intervals around the two estimated series (i.e., for twin1 and twin2) indicate that they can be reliably separated.

Fixed Interval Smoother

The Kalman filter is a recursive estimator of the (latent) state of the process at time point t, given the observations up to t. Because of the

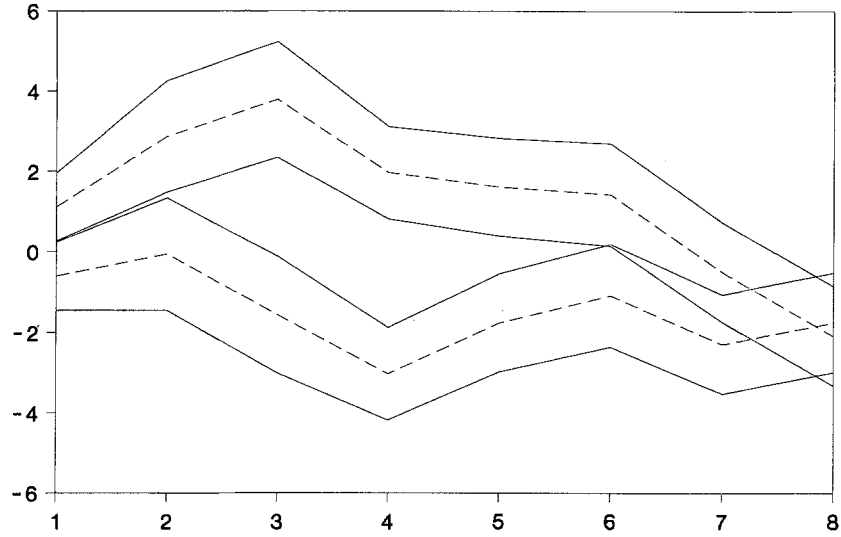

Fig. 2. Recovered genetic profiles across time for twin1 and twin2 of a DZ twin pair. The dashed lines give the recovered genetic factor scores. Solid lines represent the 95% upper and lower confidence intervals around estimated profiles.

recursive nature of the Kalman filter many computational and storage requirements are avoided that would be necessary if all data were processed at each measurement occasion. In the case of twin data, this means, for example, that for data sets consisting of large numbers of repeatedly measured variables, estimation of parameters as in Eq. (5) can be carried out on mean cross products and that these estimates can be used in the Kalman filter, thus minimizing computational and storage requirements. In cross-sectional designs the Kalman filter is identical to the well-known regression estimator of factor scores (Priestly and Subba Rao, 1975). In longitudinal designs, however, the regression estimator is optimal, while the Kalman filter is not. Optimal recursive estimates of factor scores in noisy time series can be obtained with the fixed interval smoother. The latter recursive estimates are identical to those obtained with the regression estimator. The fixed interval smoother consists of the Kalman filter followed by a recursion from the last to the first measurement occasion. During this backward sweep information is used that was obtained during Kalman filtering, *viz.*, the Kalman estimates and their respective error covariance matrices.

Two additional trivariate data sets were simulated where the vari-

ance of the noise series $\epsilon(t)$ now was equal to 100 instead of 10 for each indicator variable. For each variable more than 50% of the signal thus consisted of noise. For the first data set [$\beta(g) = \beta(e) = 0$] there were no improvements in estimates for $G(t)$ and $E(t)$ after using the smoother. For a second data set [where $\beta(g) = \beta(e) = 0.8$] standard errors of factor scores obtained in this way were smaller and the average correlation between true and estimated G rose from 0.724 to 0.797 for DZ and from 0.848 to 0.863 for MZ twins. The average correlations between true and estimated E increased from 0.700 to 0.740 for DZ and from 0.734 to 0.779 for MZ twins. Thus, there is some improvement in estimates when the fixed interval smoother is used in addition to the Kalman filter (and β is larger than zero), but this improvement is relatively small.

Filtering of Individual Data

The results reported so far are based on Kalman filtering of twin data where the observations from both members of a pair are used simultaneously in the construction of the individual factor scores. This means that a different filter is applied to MZ and DZ data. In the MZ filter, for instance, data from both twins are weighted equally in the computation of the genetic factor scores. Parameters (i.e., genetic and environmental factor loadings) that are estimated in a genetic covariance analysis of twin data or data from other genetically related individuals can also be used, however, to construct a Kalman filter that pertains to observations from a single person. Data from genetically informative samples are of course necessary to estimate heritabilities and other parameters, but actual filtering is relatively easily carried out for a single person. That is, a filter can be constructed where no data from cotwins, siblings, or other genetically related subjects are needed. For the trivariate data sets that were analyzed above, individual factor scores were recomputed, this time without using the data from the cotwin. The results show that ignoring information from cotwins affects mainly the precision with which MZ factor scores are estimated (i.e. larger standard errors and confidence intervals around estimates). This is illustrated in Fig. 3, where confidence intervals from the "twin" filter and the "individual" filter [for the last trivariate data set where $\beta(g) = 0.8$ and $\beta(e) = 0.4$] are placed around the true genetic profile of a MZ twin pair. The precision with which factor scores for DZ twins are estimated is affected to a lesser extent. Here the results from the individual filter are almost as good as when the information from the DZ cotwin is used as well.

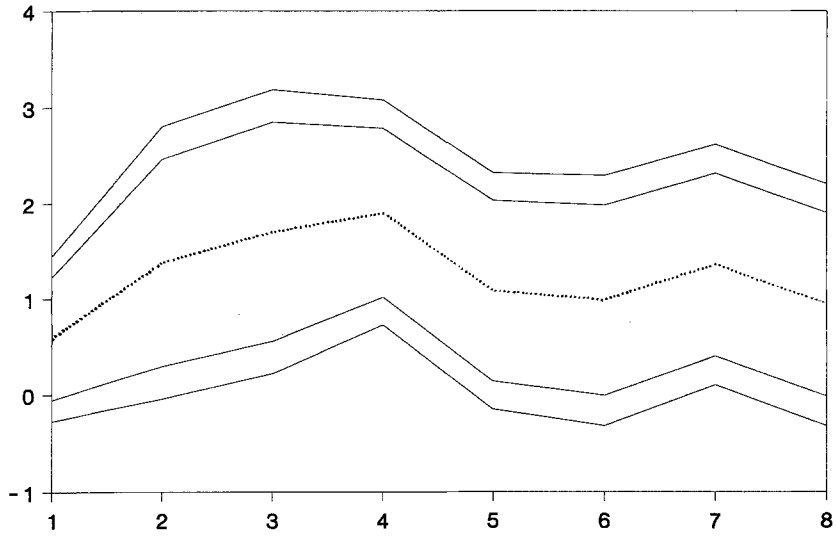

Fig. 3. Genetic profile across time for a MZ twin pair. The dashed line gives the true genetic factor scores. The inner solid lines represent the 95% confidence interval based on a Kalman filter that uses observations from both twins simultaneously. The outer solid lines represent the 95% confidence interval based on an individual Kalman filter.

DISCUSSION

The focus of this paper was the estimation of individual genetic and envionmental factor scores across time. Such factor scores can be reliably estimated and individual independent estimates of *G(t)* and *E(t)* can be obtained when at least three measures are taken at each time point. When only one measure is available, individual genetic and environmental factor scores are intercorrelated, especially for DZ twins. Results for MZ twins are quite good, however, and when bivariate measures are taken at each occasion results for both groups are acceptable. Even when there are high intercorrelations between *G(t)* and *E(t)* estimates, however, these can be taken care of by application of multivariate tests in the construction of confidence intervals (or testing differences from zero) because the Kalman filter (as the regression estimator and fixed interval smoother) yields the complete variance/covariance matrix of *G(t)* and *E(t)* estimates. Confidence intervals can also be tailored to suit specific purposes, i.e., large intervals to avoid false positives or small intervals to avoid false negatives. Estimation of *G(t)* and *E(t)* makes it possible to identify sources underlying deviant development in individual subjects. The genetic model

given by Eq. (1) can be extended to include environmental factors shared by family members and latent profiles for these processes can then be estimated. In general, a longitudinal analysis of repeated-measures data hardly allows an observed time series to be decomposed into more than one underlying series. Twin and family data are unique in that they do allow such a decomposition. Results based on a representative twin study can be generalized to the population as a whole by construction of an individual Kalman filter. Exploration of individual filtering indicates that these results are almost as good as those for DZ twins. This enables a more detailed evaluation of deviant development for individual subjects.

REFERENCES

Boomsma, D. I., and Molenaar, P. C. M. (1987). The genetic analysis of repeated measures. I. Simplex models. *Behav. Genet.* **17**:111–123.

Boomsma, D. I., Martin, N. G., and Molenaar, P. C. M. (1989). Factor and simplex models for repeated measures: Application to two psychomotor measures of alcohol sensitivity in twins. *Behav. Genet.* **19**:79–96.

Boomsma, D. I., Molenaar, P. C. M., and Orlebeke, J. F. (1990). Estimation of individual genetic and environmental factor scores. *Genet. Epidemiol.* **7**:83–91.

Brown, R. G. (1983). *Introduction to Random Signal Analysis and Kalman Filtering*, John Wiley & Sons, New York.

Eaves, L. J., Hewitt, J. K., and Heath, A. C. (1988). The quantitative genetic study of human developmental change: A model and its limitations. In Weir, B. S., Eisen, E. J., Goodman, M. M., and Namkoong, G. (eds.), *Proceedings of the Second International Conference on Quantitative Genetics*, Sinauer Associates, Sunderland, Mass., pp. 297–311.

IMSL, Inc. (1979). *IMSL Library Reference Manual Edition 7*, IMSL, Houston, Tex.

Jöreskog, K. G., and Sörbom, D. (1988). *LISREL VII, A Guide to the Program and Applications*, Spss, Chicago.

Lawley, D. N., and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*, Butterworths, London.

Martin, N. G., and Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity* **38**:79–95.

Priestley, M. B., and Subba Rao, T. (1975). The estimation of factor scores and Kalman filtering for discrete parameter stationary processes. *Int. J. Control,* **21**:971–975.

Edited by N. G. Martin